Supplementary Material for
Modeling Language Evolution Using a Spin Glass Approach

Yarahmadi, Ryom, Longobardi, Treves

This Supplementary Material provides complementary details about evolution of inter- and intra-language distances, given either the same or different random couplings.

We look at the following quantities, as they evolve in time:

S: The mean (Hamming) distance between realizations of the same language under identical coupling conditions (intra-language, identical coupling).

S': The mean distance between realizations of the same language under differing coupling conditions (intra-language, varying coupling).

D: The mean distance between realizations of different languages under identical coupling conditions (inter-language, identical coupling).

D': The mean distance between realizations of different languages under differing coupling conditions (inter-language, varying coupling).

By contrasting realizations generated under consistent versus variable conditions, we aim to disentangle the effects of shared linguistic ancestry from those due to the random assignment of couplings and to the variability in the dynamical process (i.e., in the updating sequence). Hamming distances, i.e., simply the number of differing parameters between two vectors, vary in principle in the $[0-94]$ range, but upon averaging they rarely exceed the mean value 47, or half the parameters discordant.

A Evolution of intra- and inter-language distances in the ζ–ϕ plane at $\rho = 0$

Figure S1 depicts the average (Hamming) distances between syntactic parameter vectors on the ζ-ϕ plane at $\rho = 0$, evaluated at two temporally distinct stages: at $t = 2$ (top row), i.e., in the transient regime, and at $t = 100$ (bottom row), indicating its asymptotic behavior. Distances between realizations are summarized by a color code. The columns correspond to distinct conditions: column one (Fig. S1**a**, **e**) illustrates intra-language distances with the same couplings, **S**; column two (Fig. S1**b**, **f**) presents intra-language distances with different couplings, **S'**; column three (Fig. S1**c**, **g**) inter-language distances with the same couplings, **D**; and column four (Fig. S1**d**, **h**) inter-language distances under different couplings, **D'**.

The degree of asymmetry would not appear to exert a major influence on these quantities. All average distances are lower in regions characterized by high implicational strength, and over time decrease further when computed with the same random couplings, particularly if these are close to symmetric. The main observation, however, is that all panels are prevailingly deep red, especially in the bottom right corners, where they indicate distances in the range $40 - 47$, suggesting wide dispersion of syntactic vectors. To understand more subtle effects, we need to look at differences between these measures.

The subtracted quantities, shown in Fig. S2, indicate that the major determinant of the difference in dispersal among syntactic vectors are the different sets of random couplings. Specifically, the comparisons between conditions **S'** and **S** (Fig. S2**a**, **e**) and **D'** and **D** (Fig. S2**d**, **h**) at $t = 2$ (first row) and $t = 100$ (second row), hover in the range between 5 and 20, approximately, except in the bottom right cornrs where the system is fluid, all syntactic vectors wander randomly, and then subtracted quantities approach zero. They instead reach up to about 23-24 when implications are strong, random couplings are close to symmetric and the dynamics have reached close to the asymptotic steady state (the greenish-yellow regions in the top left corners of panels Fig. S2**e** and **h**). This type of dispersal is however simply a technical aspect of the modeling approach: not knowing what the couplings between parameters could be, we assign them random values, and it is not surprising that different sets of random values lead to different steady states.

Potentially more interesting are the differences between **D'** and **S'** (Fig. S2**b**, **f**) and, even more, between **D** and **S** (Fig. S2**c**, **g**) where, in this latter case, the comparisons are obtained by keeping the same sets of

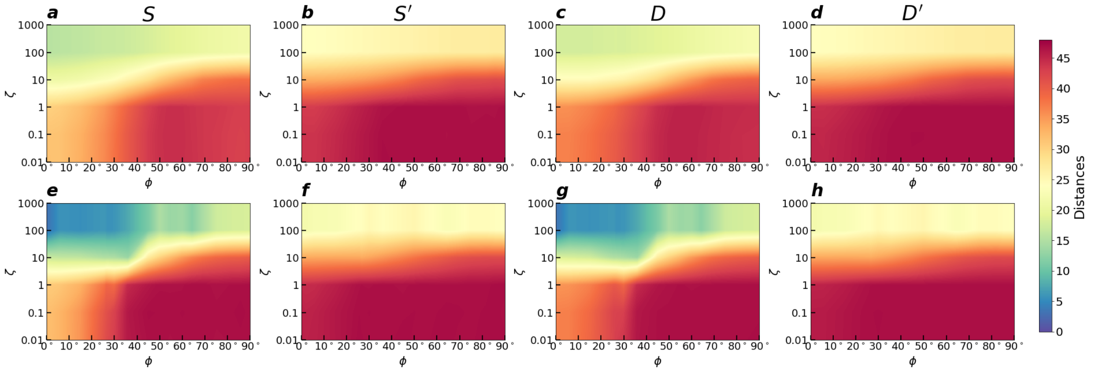

Figure S1: **Evolution of inter-language and intra-language distances in the ζ–ϕ plane at $\rho = 0$.** Hamming distances are extracted at early ($t = 2$, top row) and asymptotic ($t = 100$, bottom row) phases. (**a, e**) Average distances between realizations of same languages with identical couplings (condition **S**). (**b, f**) Average distances between realizations of same languages with different couplings (condition **S'**). (**c, g**) Average distances between realizations of the different languages with identical couplings (condition **D**). (**d, h**) Average distances between realizations of different languages with different couplings (condition **D'**). With the progression of time, all distances diminish somewhat in areas characterized by strong implicational dynamics, especially when asymmetry is minimal.

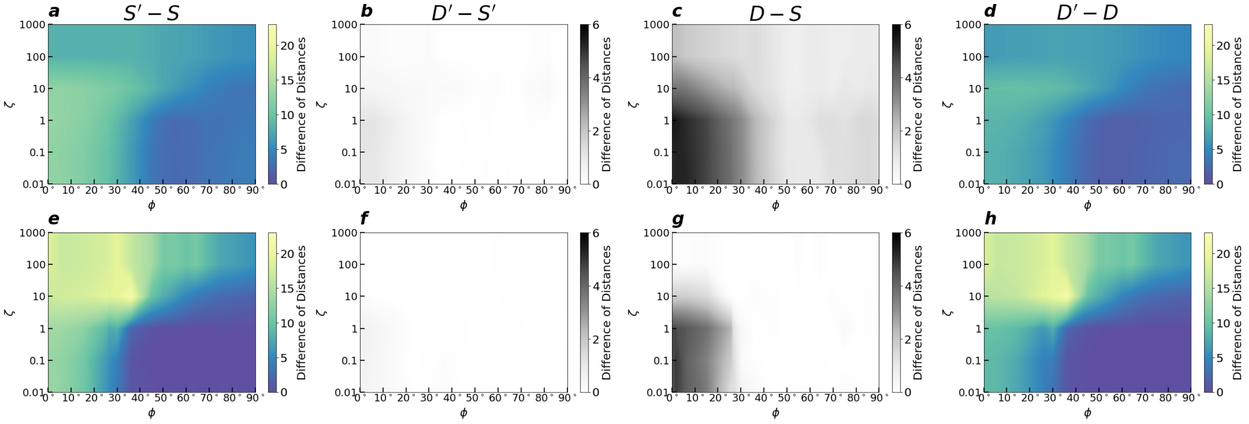

Figure S2: **Differences of distances in the ζ–ϕ plane at $\rho = 0$.** Again, distances are obtained at early ($t = 2$, top row) and asymptotic ($t = 100$, bottom row) phases. (**a, e**) Effect of coupling: difference between conditions **S'** and **S**. (**b, f**) Effect of initial conditions: difference between **D'** and **S'**, with values near zero. (**c, g**) Effect of initial conditions: difference between **D** and **S**, revealing the presence and gradual fading of the historical signal, i.e., the trace of shared ancestry among languages (see text). (**d, h**) Effect of coupling: difference between **D'** and **D**, again indicating that random coupling values generate larger differences compared to those due to initial conditions.

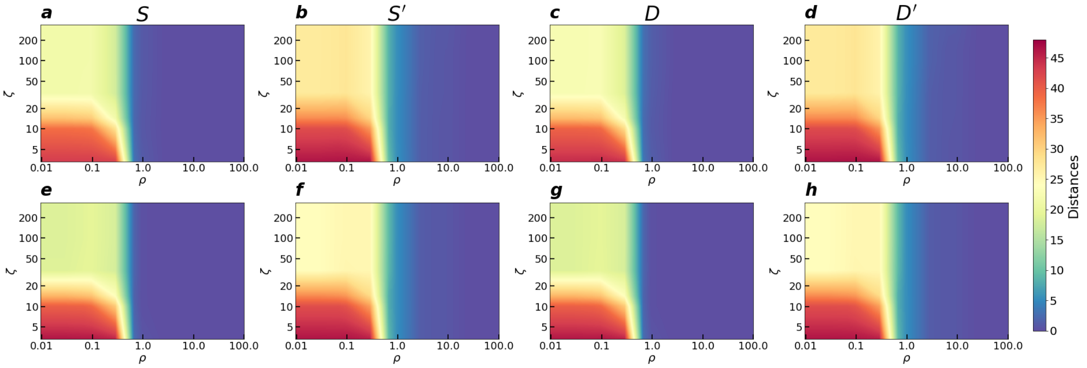

Figure S3: **Evolution of inter-language and intra-language distances in the ζ–ρ plane at $\phi = 90°$.** Computed at early ($t = 2$, top row) and asymptotic ($t = 100$, bottom row) phases. Notation and panel arrangement as in Fig. S1. Distances vanish in the regime where the Hebbian term dominates ($\rho > 1$), indicating rapid convergence toward a single shared state.

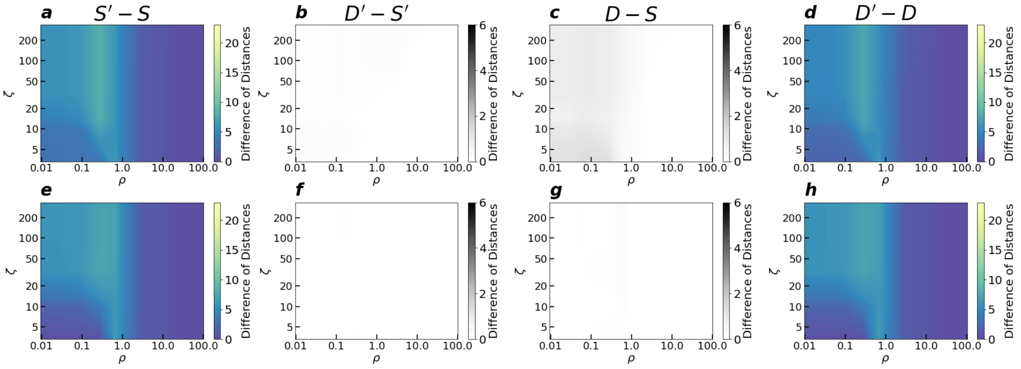

Figure S4: **Differences in Euclidean distances within the ζ–ρ plane at $\phi = 90°$.** Notation and panel arrangement as in Fig. S2.

couplings. The differences are much smaller and require us to abandon the color coding of the other panels. Fig. S2**c** and **g** quantify the phylogenetic signal, that is the differential dispersal of syntactic vectors with different ancestors, over and above that of vectors with the same ancestor. One can see that non-zero values of such a signal persist, on average, only in a region of nearly symmetric random couplings and not too strong implications. Strong implications, above $\zeta \approx 50$, are seen in our model to eventually destroy any phylogenetic signal, i.e., the genealogical relatedness of languages derived from a common ancestral source.

B Evolution of inter- and intra-language distances in the ζ–ρ plane at $\phi = 90°$

As a control, we compute the same average distances and their difference for the model extended by the inclusion of a Hebbian term, and with asymmetric couplings that counteract the Hebbian term. The very same quantities as in the section above are then calculated and plotted in the (ζ, ρ) plane, with fully antisymmetric random couplings.

Figure S3 shows that all 8 panels look similar, and indeed after just two time steps, all realizations, either corresponding to a given language or to different languages, either with the same set of random couplings or with different sets, converge onto a single trajectory when $\rho > 1$, as evidenced by blue color code. The Hebbian term makes syntactic vectors collapse onto each other. Linguistic diversity is lost when $\rho > 1$. This indicates the need of rethinking the Hebbian term and possibly adopting a more sophisticated form, which

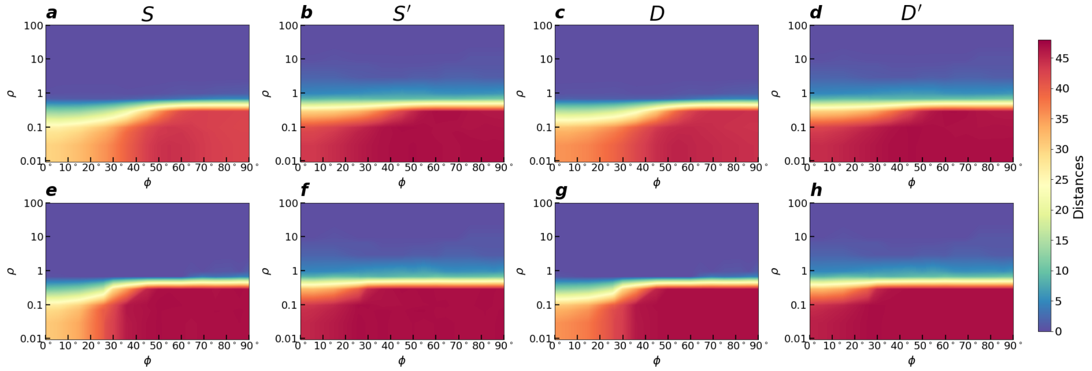

Figure S5: **Changes in intra-language and inter-language distances on a the ρ–ϕ plane with $\zeta = 0.01$.** Computed at an early ($t = 2$, top row) and a long-term ($t = 100$, bottom row) phase. (Notation and panel arrangement as in Fig. S1. All distance vanish when the Hebbian term dominates ($\rho > 1$), reflecting rapid collapse onto a single state.

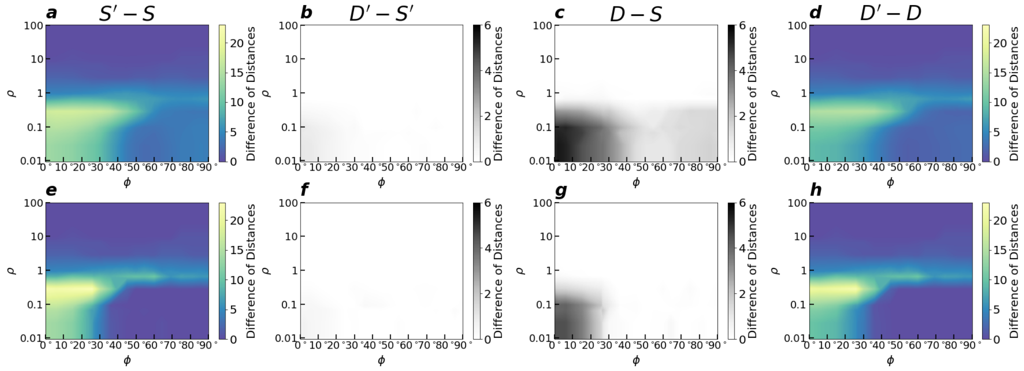

Figure S6: **Differences in distances on the ρ–ϕ plane at $\zeta = 0.01$.** Notation and panel arrangement as in Fig. S2.

does not lead to the collapse.

The differences between distance measures, shown in Fig. S4, indicate that now random coupling effects are much reduced. Notably, when subtracting distances obtained with the same random couplings from those obtained with different ones (Fig. S4**a**, **d**, **e** and **h**) the subtracted values range below about 10, and below 5 is $\rho > 1$. In contrast, comparisons between vectors evolving from the same or different ancestors, either across coupling sets (Fig. S4**b**, **f**) or within coupling sets (Fig. S4**c**, **g**), which isolate the effect of initial conditions - the phylogenetic signal - yield values < 1. This further supports the conclusion that strong asymmetry and a robust Hebbian term (in its present form) are incompatible with the observation of a historical signal.

C Evolution of inter- and intra-language distances in the ρ–ϕ plane at $\zeta = 0.01$

Finally, for completeness we look at the same average distances and their difference on a third orthogonal plane, defined by ϕ and ρ, with the strength of the implications fixed at the low value $\zeta = 0.01$. The same quantities as in the two sections above are then calculated and plotted as a function of (ϕ, ρ) plane, with weak implications.

Figure S5 is again comprised of 8 similar panels, all showing that syntactic vectors always collapse onto each other, whatever their origin, when the strength of the Hebbian term is sufficient, e.g. $\rho > 1$. Below this regime, on the right where fluid dynamics prevail all phylogenetic relations are rapidly washed away and distances tend to their random values. The better appreciate what happens in the bottom left corners, we turn again to the difference plots.

The differences in Fig. S6 point at the effect of changing the random couplings, largely confined to the low-ρ, symmetric regime but also extending along the border between the fluid and the high-ρ regime (Fig. S6**a**, **d**), **e** and **h**). Interestingly, Fig. S6**c**, **g** show that the historical signal, still present at $t = 100$ in the bottom left corner of symmetric couplings and frozen dynamics, at an early stage, $t = 2$ is visible also in the fluid dynamics regime to the right.

Overall, the analysis of intra- and inter-language distances indicates that, in this model, the past leaves its trace, in syntax, only if syntax evolves constrained by nearly symmetric couplings, with weak (asymmetric) implications and weak (symmetric) Hebbian terms. Understanding whether these conclusions apply in a wider class of models, getting closer to linguistic plausibility, requires further work.